

# Soft-Label Dataset Distillation and Text Dataset Distillation

**Ilia Sucholutsky**

*Department of Statistics and Actuarial Science  
University of Waterloo  
Waterloo, Ontario, Canada*

ISUCHOLU@UWATERLOO.CA

**Matthias Schonlau**

*Department of Statistics and Actuarial Science  
University of Waterloo  
Waterloo, Ontario, Canada*

SCHONLAU@UWATERLOO.CA

**Editor:**

## Abstract

Dataset distillation is a method for reducing dataset sizes by learning a small number of synthetic samples containing all the information of a large dataset. This has several benefits like speeding up model training, reducing energy consumption, and reducing required storage space. Currently, each synthetic sample is assigned a single ‘hard’ label, and also, dataset distillation can currently only be used with image data.

We propose to simultaneously distill both images and their labels, thus assigning each synthetic sample a ‘soft’ label (a distribution of labels). Our algorithm increases accuracy by 2-4% over the original algorithm for several image classification tasks. Using ‘soft’ labels also enables distilled datasets to consist of fewer samples than there are classes as each sample can encode information for multiple classes. For example, training a LeNet model with 10 distilled images (one per class) results in over 96% accuracy on MNIST, and almost 92% accuracy when trained on just 5 distilled images.

We also extend the dataset distillation algorithm to distill sequential datasets including texts. We demonstrate that text distillation outperforms other methods across multiple datasets. For example, models attain almost their original accuracy on the IMDB sentiment analysis task using just 20 distilled sentences.

Our code can be found at <https://github.com/ilia10000/dataset-distillation>

**Keywords:** Dataset Distillation, Knowledge Distillation, Neural Networks, Synthetic Data, Gradient Descent

## 1. Introduction

The increase in computational requirements for modern deep learning presents a range of issues. The training of deep learning models has an extremely high energy consumption (Strubell et al., 2019), on top of the already problematic financial cost and time requirement. One path for mitigating these issues is to reduce network sizes. Hinton et al. (2015) proposed knowledge distillation as a method for imbuing smaller, more efficient networks with all the knowledge of their larger counterparts. Instead of decreasing network size, a second path to efficiency may be to decrease dataset size. Dataset distillation (DD)

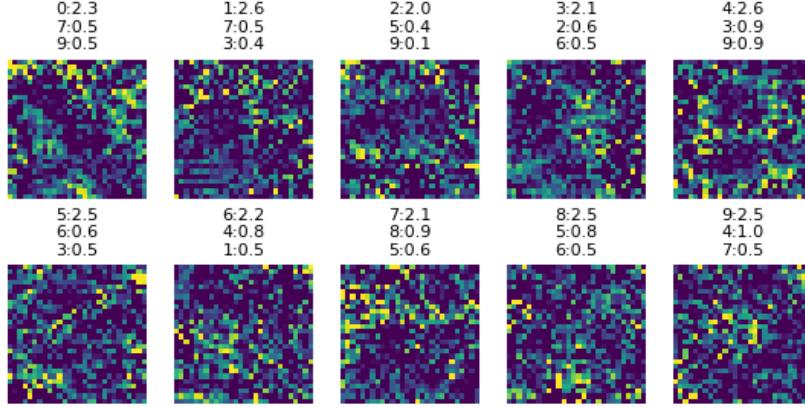


Figure 1: 10 MNIST images learned by SLDD can train networks with fixed initializations from 11.13% distillation accuracy to 96.13% ( $r_{10} = 97.1$ ). Each image is labeled with its top 3 classes and their associated logits. The full labels for these 10 images are in Table 1.

Table 1: Learned distilled labels for the 10 distilled MNIST images in Figure 1. Distilled labels are allowed to take on any real value. If a probability distribution is needed, a softmax function can be applied to each row.

Distilled Label	Digit									
	0	1	2	3	4	5	6	7	8	9
1	2.34	-0.33	0.23	0.04	-0.03	-0.23	-0.32	0.54	-0.39	0.49
2	-0.17	2.58	0.32	0.37	-0.68	-0.19	-0.75	0.53	0.27	-0.89
3	-0.26	-0.35	2.00	0.07	0.08	0.42	0.02	-0.08	-1.09	0.10
4	-0.28	0.04	0.59	2.08	-0.61	-1.11	0.52	0.19	-0.20	0.32
5	-0.11	-0.52	-0.08	0.90	2.63	-0.44	-0.72	-0.39	-0.29	0.87
6	0.25	-0.20	-0.19	0.51	-0.02	2.47	0.62	-0.42	-0.52	-0.63
7	0.42	0.55	-0.09	-1.07	0.83	-0.19	2.16	-0.30	0.26	-0.91
8	0.18	-0.33	-0.25	0.06	-0.91	0.55	-1.17	2.11	0.94	0.47
9	0.46	-0.48	0.24	0.09	-0.78	0.75	0.47	-0.40	2.45	-0.71
10	-0.53	0.52	-0.74	-1.32	1.03	0.23	0.05	0.55	0.31	2.45

has recently been proposed as an alternative formulation of knowledge distillation to do exactly that (Wang et al., 2018).

Dataset distillation is the process of creating a small number of synthetic samples that can quickly train a network to the same accuracy it would achieve if trained on the original dataset. It may seem counter-intuitive that training a model on a small number of synthetic images coming from a different distribution than the training data can result in comparable accuracy, but Wang et al. (2018) have shown that for models with known initializations this is indeed feasible; they achieve 94% accuracy on MNIST, a hand-written digit recognition task (LeCun et al., 1998), after training LeNet on just 10 synthetic images. We propose to

Figure 2: **Left:** An example of a ‘hard’ label where the second class is selected. **Center:** An example of a ‘soft’ label restricted to being a valid probability distribution. The second class has the highest probability. **Right:** An example of an unrestricted ‘soft’ label. The second class has the highest weight. ‘Hard’ labels can be derived from unrestricted ‘soft’ labels by applying the softmax function and then setting the highest probability element to 1, and the rest to 0.

$$\begin{array}{c|c|c}
 \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} & 
 \begin{bmatrix} 0.01 \\ 0.69 \\ 0.02 \\ 0.02 \\ 0.03 \\ 0.05 \\ 0.03 \\ 0.1 \\ 0.01 \\ 0.04 \end{bmatrix} & 
 \begin{bmatrix} 0.8 \\ 5.1 \\ 1.5 \\ 1.5 \\ 2 \\ 2.5 \\ 2 \\ 3.2 \\ 0.8 \\ 2 \end{bmatrix} \\
 \xleftarrow[\text{and rest to 0}]{\text{Set largest to 1}} & 
 \xleftarrow[\text{softmax}]{\text{Apply}} & 
 \end{array}$$

improve their already impressive results by learning ‘soft’ labels as a part of the distillation process. The original dataset distillation algorithm uses fixed, or ‘hard’, labels for the synthetic samples (e.g. the ten synthetic MNIST images each have a label corresponding to a different digit). In other words, each label is a one-hot vector: a vector where all entries are set to zero aside from a single entry, the one corresponding to the correct class, which is set to one. We relax this one-hot restriction and make the synthetic labels learnable. The resulting distilled labels are thus similar to those used for knowledge distillation as a single image can now correspond to multiple classes. An example comparing a ‘hard’ label to a ‘soft’ label is shown in Figure 2. A ‘hard’ label can be derived from a ‘soft’ label by applying the softmax function and setting the element with the highest probability to one, while the remaining elements are set to zero. Our soft-label dataset distillation (SLDD) not only achieves over 96% accuracy on MNIST when using ten distilled images (as seen in Figure 1), a 2% increase over the state-of-the-art (SOTA), but also achieves almost 92% accuracy with just five distilled images, which is less than one image per class. In addition to soft labels, we also extend dataset distillation to the natural language/sequence modeling domain and enable it to be used with several additional neural network architectures. For example, we show that Text Dataset Distillation (TDD) can train a custom convolutional neural network (CNN) (LeCun et al., 1999) with known initialization up to 90% of its original accuracy on the IMDB sentiment classification task (Maas et al., 2011) using just two synthetic sentences.

The rest of this work is divided into four sections. In Section 2, we discuss related work in the fields of knowledge distillation, dataset reduction, and example generation. In Section 3, we propose improvements and extensions to dataset distillation and associated theory. In Section 4, we empirically validate SLDD and TDD in a wide range of experiments. Finally, in Section 5, we discuss the significance of SLDD and TDD, and our outlook for the future.

## 2. Related Work

### 2.1 Knowledge Distillation

Dataset distillation was originally inspired by network distillation (Hinton et al., 2015) which is a form of knowledge distillation or model compression (Bucilu et al., 2006). Network distillation has been studied in various contexts including when working with sequential data (Kim and Rush, 2016). Network distillation aims to distill the knowledge of large, or even multiple, networks into a smaller network. Similarly, dataset distillation aims to distill the knowledge of large, or even multiple, datasets into a small number of synthetic samples. ‘Soft’ labels were recently proposed as an effective way of distilling networks by feeding the output probabilities of a larger network directly to a smaller network (Hinton et al., 2015), and have previously been studied in the context of different machine learning algorithms (El Gayar et al., 2006). Our soft-label dataset distillation (SLDD) algorithm also uses ‘soft’ labels but these are persistent and learned over the training phase of a network (rather than being produced during the inference phase as in the case of network distillation).

### 2.2 Learning from ‘small’ data

Deep supervised learning generally requires a very large number of examples to train on. For example, MNIST and CIFAR10 both contain thousands of training images per class. Meanwhile, it appears that humans can quickly generalize from a tiny number of examples (Lake et al., 2015). Getting machines to learn from ‘small’ data is an important aspect of trying to bridge this gap in abilities. Dataset distillation provides a method for researchers to generate synthetic examples that are optimized for allowing machines to learn from a small number of them. Studying the distilled images produced by dataset distillation may enable us to identify what allows neural networks to generalize so quickly from so few of them. In some sense, dataset distillation can be thought of as an algorithm for creating dataset summaries that machines can learn from.

### 2.3 Dataset Reduction, Prototype Generation, and Summarization

There are a large number of methods that aim to reduce the size of a dataset with varying objectives. Active learning aims to reduce the required size of the labeled portion of a dataset by only labeling examples that are determined to be the most important (Cohn et al., 1996; Tong and Koller, 2001). Several methods aim to ‘prune’ a dataset, or create a ‘core-set’, by leaving in only examples that are determined to be useful (Angelova et al., 2005; Bachem et al., 2017; Sener and Savarese, 2017; Tsang et al., 2005). In general, all of these methods use samples from the true distribution, typically subsets of the original training set. By lifting this restriction and, instead, learning synthetic samples, dataset distillation requires far fewer samples to distill the same amount of knowledge.

In the field of nearest-neighbor classification, these dataset reduction techniques are typically referred to as ‘prototype selection’ and ‘prototype generation’, and are studied extensively as methods of reducing storage requirements and improving the efficiency of nearest-neighbor classification (Garcia et al., 2012; Triguero et al., 2011). As with the methods above, prototype selection methods use samples from the true distribution, typi-

cally just subsets of the original training dataset. Prototype generation methods typically create samples that are not found in the training data; however, these methods are designed specifically for use with nearest-neighbor classification algorithms.

All the dataset reduction methods discussed above also share another restriction. They all use fixed labels. Soft-label dataset distillation removes this restriction and allows the label distribution to be optimized simultaneously with the samples (or prototypes) themselves.

## 2.4 Generative Adversarial Networks

Generative Adversarial Networks (GANs) have recently become a widely used method for image generation. They are primarily used to produce images that closely mimic those coming from the true distribution (Ledig et al., 2017; Goodfellow et al., 2014; Choi et al., 2018; Radford et al., 2015). For dataset distillation, we instead set knowledge distillation as the objective but do not attempt to produce samples from the true distribution. Using the generator from a trained GAN may be a much faster way of producing images than the gradient-based method employed by dataset distillation. However, since the number of distilled images we aim to produce is very small, solving the objective directly through gradient-based optimization is sufficiently fast, while also more straightforward. Additionally, while some GANs can work with text (Reed et al., 2016; Yu et al., 2017), they are primarily intended for image generation.

## 2.5 Measuring Problem Dimensionality

We may intuitively believe that one deep learning task is more difficult than another. For example, when comparing the digit recognition task MNIST, to the image classification task CIFAR10 (Krizhevsky et al., 2009), CIFAR10 appears to be the more difficult problem though it is difficult to measure the extent of this increase in difficulty. One approach is to compare error rates for state-of-the-art (SOTA) results on these datasets. For example, the near-SOTA ‘dropconnect’ model on MNIST achieves a 0.21% error rate, while on CIFAR10 it achieves an error rate of 9.32% (Wan et al., 2013). However, this approach reveals less as deeper networks approach perfect accuracy on multiple tasks. Li et al. (2018) instead derive a fairly model-independent metric for comparing the dimensionality of various problems based on the minimum number of learnable parameters needed to achieve a good local optimum. Similarly, dataset distillation aims to find the minimum number of synthetic samples needed to achieve a good local optimum. The difference is that Li et al. (2018) constrain the number of searchable dimensions within the network weight space, while dataset distillation constrains them within the data space.

### 3. Extending Dataset Distillation

#### 3.1 Motivation

As mentioned above, nearest-neighbors classification often involves data reduction techniques known as prototype selection and prototype generation. We can use these concepts, along with the k-Nearest Neighbors (kNN) classification algorithm, to visualize the difference between classical dataset reduction methods, dataset distillation, and soft-label dataset distillation. When we fit a kNN model, we essentially divide the entire space into classes based on the location of points in the training set. However, the cost of fitting a kNN model increases with the number of training points. Prototype selection and generation are methods for reducing the number of points in the training set while trying to maintain the accuracy of the original model.

Prototype selection methods use subsets of the training set to construct the reduced training set. In the first column of Figure 3, we visualize attempts to reduce the training set for a three-class problem, the Iris flower dataset (Fisher, 1936), by selecting one point from each class and then fitting the kNN on these three selected points. It is clear from this visualization that only being able to use a subset of the original points, limits the ability to finely tune the resulting separation of the space. Prototype generation methods relax this restriction and create synthetic points whose placement can be optimized for resulting kNN performance. This is effectively what dataset distillation does for neural networks. In the second column of Figure 3, we generate one point for each class and then optimize the location of one of these points, before fitting the kNN on all three. Using synthetic, optimizable points allows for finer-grained tuning of the resulting class separation. In both these cases, each of the three resulting points had a single class assigned to it.

We now propose that the three points instead be assigned an optimizable distribution of classes, effectively a ‘soft’ label as described above. In the third column of Figure 3, in order to visualize the effect of changing a point’s label distribution, we arbitrarily fix one point for each class, but we change the label distribution of one of these points, increasingly making it a mixture of the other classes, and then fit the kNN. In the final column of Figure 3, we combine the prototype generation method with our soft-label modification, to visualize the effect of simultaneously changing a point’s location and label distribution. This last case is the kNN counterpart to our proposed soft-label distillation algorithm, and from the visualization, it is clear that it provides the finest-grained tuning for the resulting class separation. In fact, by using soft labels with kNN, we can separate three classes using just two points, as seen in Figure 4.

Animated versions of both Figure 3 and Figure 4 are in the online appendix.

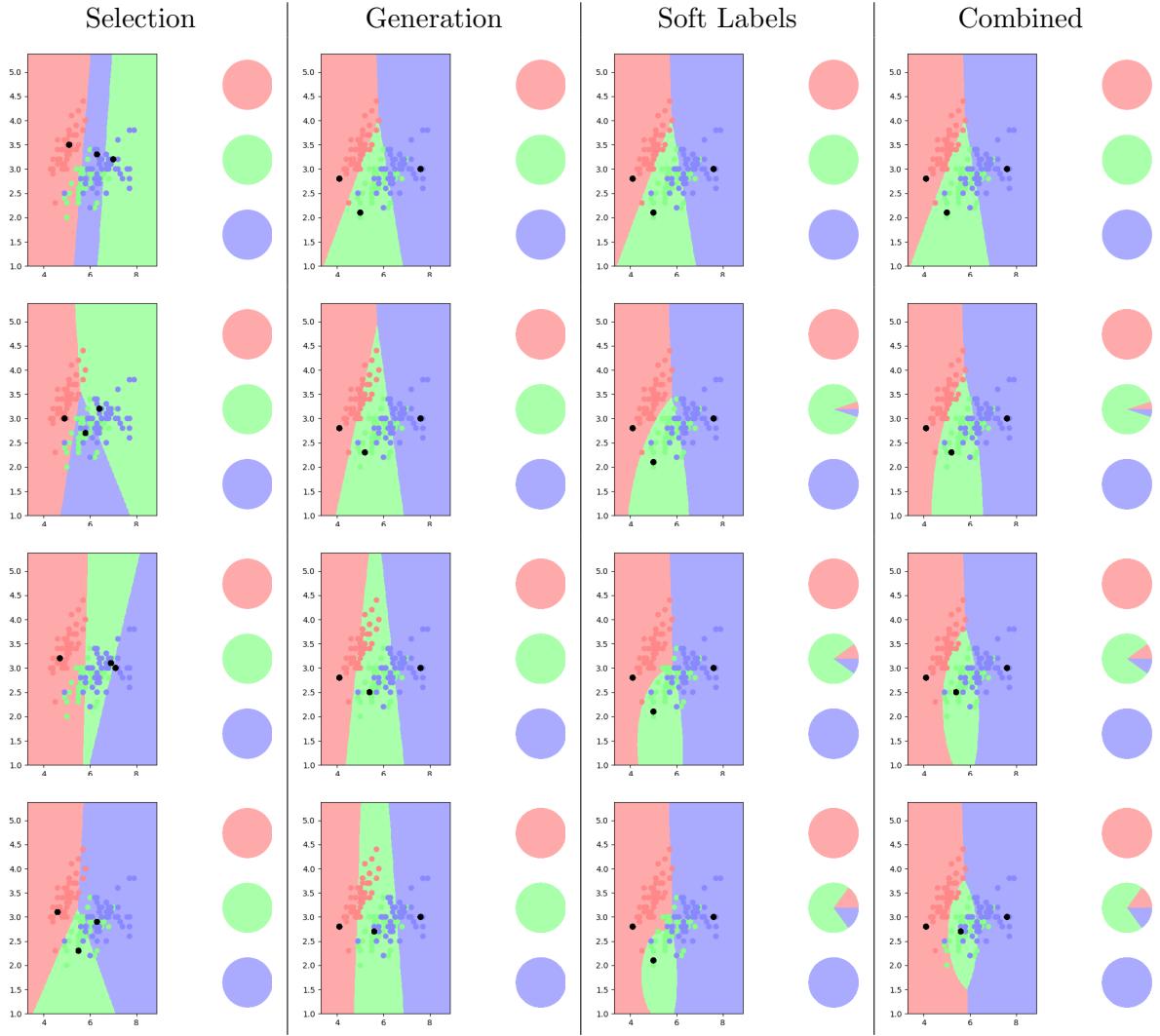


Figure 3: kNN models are fitted on 3 points obtained from the Iris flower dataset using four methods: prototype selection, prototype generation, soft labels, and prototype generation combined with soft labels. Each column contains 4 steps of the associated method used to update the 3 points used to fit the associated kNN. The pie charts represent the label distributions assigned to each of the 3 points.

**Selection method:** A different random point from each class is chosen to represent its class in each of the steps.

**Generation method:** The middle point associated with the 'green' label is moved diagonally in each step.

**Soft labels method:** The label distribution of the middle point is changed each step to contain a larger proportion of both other classes.

**Combined method:** The middle point is simultaneously moved and has its label distribution updated in each step.

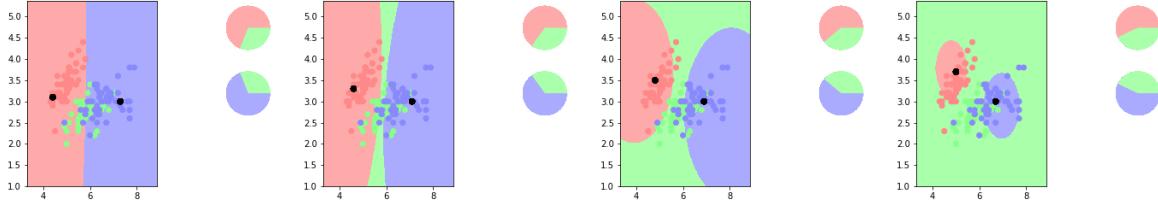


Figure 4: kNN model fitted on 2 points obtained using a combination of prototype generation and soft labels. The pie charts represent the label distributions assigned to each of the 2 points. From left-to-right, in each plot, the locations of the 2 points are slightly shifted and the values associated with their ‘green’ label are increased. By modifying the location and soft labels of the 2 points, the space can still be separated into 3 classes.

### 3.2 Basic Approach

Our basic approach is the same as Wang et al. (2018). We summarize it here in a slightly modified way to explicitly show the labels of the distilled dataset. This additional notation becomes useful once we enable label learning in the next section.

Given a training dataset  $\mathbf{d} = \{x_i, y_i\}_{i=1}^N$ , a neural network with parameters  $\theta$ , and a twice-differentiable loss function  $\ell(x_i, y_i, \theta)$ , our objective is to find

$$\theta^* = \arg \min_{\theta} \frac{1}{N} \sum_{i=1}^N \ell(x_i, y_i, \theta) \triangleq \arg \min_{\theta} \ell(\mathbf{x}, \mathbf{y}, \theta). \quad (1)$$

In general, training with stochastic gradient descent (SGD) involves repeatedly sampling mini-batches of training data and updating network parameters by their error gradient scaled by learning rate  $\eta$ .

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta_t} \ell(\mathbf{x}_t, \mathbf{y}_t, \theta_t) \quad (2)$$

With dataset distillation, the goal is to perform just one such step while still achieving the same accuracy. We do this by learning a very small number of synthetic samples  $\tilde{\mathbf{x}}$  that minimize  $\mathcal{L}$ , a one-step loss objective, for  $\theta_1 = \theta_0 - \tilde{\eta} \nabla_{\theta_0} \ell(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \theta_0)$ .

$$\mathcal{L}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\eta}; \theta_0) := \ell(\mathbf{x}, \mathbf{y}, \theta_1) = \ell(\mathbf{x}, \mathbf{y}, \theta_0 - \tilde{\eta} \nabla_{\theta_0} \ell(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \theta_0)) \quad (3)$$

$$\tilde{\mathbf{x}}^*, \tilde{\eta}^* = \arg \min_{\tilde{\mathbf{x}}, \tilde{\eta}} \mathcal{L}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\eta}; \theta_0) = \arg \min_{\tilde{\mathbf{x}}, \tilde{\eta}} \ell(\mathbf{x}, \mathbf{y}, \theta_0 - \tilde{\eta} \nabla_{\theta_0} \ell(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \theta_0)) \quad (4)$$

Note that, currently, we are minimizing over  $\tilde{\mathbf{x}}$  and  $\tilde{\eta}$ , but not  $\tilde{\mathbf{y}}$ , as the distilled labels are fixed for the original dataset distillation algorithm. We minimize this objective, or in other words ‘learn the distilled samples’, by using standard gradient descent.

### 3.3 Learnable Labels

As mentioned above, one formulation of knowledge distillation proposes that a smaller network be trained on the outputs of a larger network rather than the original training labels. Unlike the training labels, the output labels of the larger network are not ‘hard’ labels. Because they are generally the outputs of a softmax layer, the output labels form a probability distribution over the possible classes. The idea is that any training image contains information about more than one class (e.g. an image of the digit ‘3’ looks a lot like other digits ‘3’ but it also looks like the digit ‘8’). Using ‘soft’ labels allows us to convey more information about the associated image.

The original dataset distillation algorithm was restricted to ‘hard’ labels for the distilled data; each distilled image has to be associated with just a single class. We relax this restriction and allow distilled labels to take on any real value. Since the distilled labels are now continuous variables, we can modify the distillation algorithm in order to make the distilled labels learnable using the same method as for the distilled images: a combination of backpropagation and gradient descent. With our modified notation, we simply need to change equation (4) to also minimize over  $\tilde{\mathbf{y}}$ .

$$\tilde{\mathbf{x}}^*, \tilde{\mathbf{y}}^*, \tilde{\eta}^* = \arg \min_{\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\eta}} \mathcal{L}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\eta}; \theta_0) = \arg \min_{\tilde{\mathbf{x}}, \tilde{\eta}} \ell(\mathbf{x}, \mathbf{y}, \theta_0 - \tilde{\eta} \nabla_{\theta_0} \ell(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \theta_0)) \quad (5)$$

Algorithm 1a details this soft-label dataset distillation (SLDD) algorithm. We note that in our experiments, we generally initialize  $\tilde{y}$  with the one-hot values that ‘hard’ labels would have. We found that this tends to increase accuracy when compared to random initialization, perhaps because it encourages more differentiation between classes early on in the distillation process.

### 3.4 Text and Other Sequences

The original dataset distillation algorithm was only shown to work with image data, but intuitively, there is no reason why text or other sequences should not be similarly distillable. However, it is difficult to use gradient methods directly on text data as it is discrete. In order to be able to use SLDD with text data, we need to first embed the text data into a continuous space. This is a common practice when working with many modern natural language processing models, though the embedding method itself can vary greatly (Ma and Hovy, 2016; Devlin et al., 2018; Peters et al., 2018). Any popular embedding method can be used; in our experiments, we used pre-trained GloVe embeddings (Pennington et al., 2014). Once the text is embedded into a continuous space, the problem of distilling it becomes analogous to soft-label image distillation. If all sentences are padded/truncated to some pre-determined length, then each sentence is essentially just a one-channel image of size [length]\*[embedding dimension]. When working with models that do not require fixed-length input, like recurrent neural networks (RNN), the distilled sentences do not have to be padded/truncated to the same length. It is also important to note that the embedding is performed only on sentences coming from the true dataset; the distilled samples are learned directly as embedded representations. Since the distilled data produced by this algorithm would still be in the embedding space, it may be of interest to find the nearest sentences that correspond to the distilled embeddings. To compute the nearest sentence to

---

**Algorithm 1a** Soft-Label Dataset Distillation (SLDD)

---

**Input:**  $p(\theta_0)$ : distribution of initial weights;  $M$ : the number of distilled data;  $\alpha$ : step size;  $n$ : batch size;  $T$ : number of optimization iterations;  $\tilde{y}_0$ : initial value for  $\tilde{y}$ ;  $\tilde{\eta}_0$ : initial value for  $\tilde{\eta}$

- 1: Initialize distilled data  
 $\tilde{\mathbf{x}} = \{\tilde{x}_i\}_{i=1}^M$  randomly,  
 $\tilde{\mathbf{y}} = \{\tilde{y}_i\}_{i=1}^M \leftarrow \tilde{y}_0$ ,  
 $\tilde{\eta} \leftarrow \tilde{\eta}_0$
- 2: **for** each training step  $t = 1$  to  $T$  **do**
- 3:   Get a mini-batch of real training data  
 $(\mathbf{x}_t, \mathbf{y}_t) = \{x_{t,j}, y_{t,j}\}_{j=1}^n$
- 4:   One-hot encode the labels  
 $(\mathbf{x}_t, \mathbf{y}^*_t) = \{x_{t,j}, \text{Encode}(y_{t,j})\}_{j=1}^n$
- 5:   Sample a batch of initial weights  
 $\theta_0^{(j)} \sim p(\theta_0)$
- 6:   **for** each sampled  $\theta_0^{(j)}$  **do**
- 7:     Compute updated model parameter with GD  
 $\theta_1^{(j)} = \theta_0^{(j)} - \tilde{\eta} \nabla_{\theta_0^{(j)}} \ell(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \theta_0^{(j)})$
- 8:     Evaluate the objective function on real training data:  $\mathcal{L}^{(j)} = \ell(\mathbf{x}_t, \mathbf{y}^*_t, \theta_1^{(j)})$
- 9:   **end for**
- 10:   Update distilled data  
 $\tilde{\mathbf{x}} \leftarrow \tilde{\mathbf{x}} - \alpha \nabla_{\tilde{\mathbf{x}}} \sum_j \mathcal{L}^{(j)}$ ,  
 $\tilde{\mathbf{y}} \leftarrow \tilde{\mathbf{y}} - \alpha \nabla_{\tilde{\mathbf{y}}} \sum_j \mathcal{L}^{(j)}$ , and  
 $\tilde{\eta} \leftarrow \tilde{\eta} - \alpha \nabla_{\tilde{\eta}} \sum_j \mathcal{L}^{(j)}$
- 11: **end for**

**Output:** distilled data  $\tilde{\mathbf{x}}$ ; distilled labels  $\tilde{\mathbf{y}}$ ; optimized learning rate  $\tilde{\eta}$

---

a distilled embedding matrix, for every column vector in the matrix, the nearest embedding vector from the original dictionary must be found. These embedding vectors must then be converted back into their corresponding words, and those words joined into a sentence. The resulting algorithm for text dataset distillation (TDD) is detailed in Algorithm 1b which is a modification of the SLDD Algorithm 1a.

### 3.5 Random initializations and multiple steps

The procedures we described above make one important assumption: network initialization  $\theta_0$  is fixed. The samples created this way do not lead to high accuracies when the network is re-trained on them with a different initialization as they contain information not only about the dataset but also about  $\theta_0$ . In the distilled images in Figures 1 and 5, this can be seen as what looks like a lot of random noise. Wang et al. (2018) propose that the method instead be generalized to work with network initializations randomly sampled from some restricted distribution.

$$\tilde{\mathbf{x}}^*, \tilde{\mathbf{y}}^*, \tilde{\eta}^* = \arg \min_{\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\eta}} \mathbb{E}_{\theta_0 \sim p(\theta_0)} \mathcal{L}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\eta}; \theta_0) \quad (6)$$

The resulting images, especially for MNIST, appear to have much clearer patterns and much less random noise, and the results detailed in Section 4 suggest that this method generalizes fairly well to other randomly sampled initializations from the same distribution.

Additionally, Wang et al. (2018) suggest that the above methods can work with multiple gradient descent (GD) steps. If we want to perform multiple gradient descent steps, each with a different mini-batch of distilled data, we simply backpropagate the gradient through every one of these additional steps. Finally, it may also be beneficial to train the neural networks on the distilled data for more than one epoch. The experimental results suggest that multiple steps and multiple epochs improve distillation performance for both image and text data, particularly when using random network initializations.

## 4. Experiments

### 4.1 Metrics

The simplest metric for gauging distillation performance is to train a model on distilled samples and then test it on real samples. We refer to the accuracy achieved on these real samples as the ‘distillation accuracy’. However, several of the models we use in our experiments do not achieve SOTA accuracy on the datasets they are paired with, so it is useful to construct a relative metric that compares distillation accuracy to original accuracy. The first such metric is the ‘distillation ratio’ which we define as the ratio of distillation accuracy to original accuracy. The distillation ratio is heavily dependent on the number of distilled samples so the notation we use is  $r_M = 100\% * \frac{[\text{distillation accuracy}]}{[\text{original accuracy}]}$ ,  $M = [\text{number of distilled samples}]$ . We may refer to this metric as the ‘ $M$ -sample distillation ratio’ when clarification is needed. It may also be of interest to find the minimum number of distilled images required to achieve a certain distillation ratio. To this end we can define a second relative metric that we call the ‘ $A\%$  distillation size’, and we write  $d_A = M$  where  $M$  is the minimum number of distilled samples required to achieve a distillation ratio of  $A\%$ .

---

**Algorithm 1b** Text Dataset Distillation (TDD)

---

**Input:**  $p(\theta_0)$ : distribution of initial weights;  $M$ : the number of distilled data;  $\alpha$ : step size;  $n$ : batch size;  $T$ : number of optimization iterations;  $\tilde{y}_0$ : initial value for  $\tilde{y}$ ;  $\tilde{\eta}_0$ : initial value for  $\tilde{\eta}$ ;  $s$ : sentence length;  $d$ : embedding size

- 1: Initialize distilled data  
 $\tilde{\mathbf{x}} = \{\tilde{x}_i\}_{i=1}^M$  randomly of size  $s \times d$ ,  
 $\tilde{\mathbf{y}} = \{\tilde{y}_i\}_{i=1}^M \leftarrow \tilde{y}_0$ ,  
 $\tilde{\eta} \leftarrow \tilde{\eta}_0$
  - 2: **for** each training step  $t = 1$  to  $T$  **do**
  - 3:   Get a mini-batch of real training data  
 $(\mathbf{x}_t, \mathbf{y}_t) = \{x_{t,j}, y_{t,j}\}_{j=1}^n$
  - 4:   Pad (or truncate) each sentence in the mini-batch  
 $(\mathbf{x}^p_t, \mathbf{y}_t) = \{\text{Pad}(x_{t,j}, \text{len} = s), y_{t,j}\}_{j=1}^n$
  - 5:   Embed each sentence in the mini-batch  
 $(\mathbf{x}^*_t, \mathbf{y}_t) = \left\{ \text{Embed}(x_{t,j}^p, \text{dim} = d), y_{t,j} \right\}_{j=1}^n$
  - 6:   One-hot encode the labels  
 $(\mathbf{x}^*_t, \mathbf{y}^*_t) = \left\{ x_{t,j}^*, \text{Encode}(y_{t,j}) \right\}_{j=1}^n$
  - 7:   Sample a batch of initial weights  
 $\theta_0^{(j)} \sim p(\theta_0)$
  - 8:   **for** each sampled  $\theta_0^{(j)}$  **do**
  - 9:     Compute updated model parameter with GD  
 $\theta_1^{(j)} = \theta_0^{(j)} - \tilde{\eta} \nabla_{\theta_0^{(j)}} \ell(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \theta_0^{(j)})$
  - 10:    Evaluate the objective function on real training data:  $\mathcal{L}^{(j)} = \ell(\mathbf{x}^*_t, \mathbf{y}^*_t, \theta_1^{(j)})$
  - 11:   **end for**
  - 12:   Update distilled data  
 $\tilde{\mathbf{x}} \leftarrow \tilde{\mathbf{x}} - \alpha \nabla_{\tilde{\mathbf{x}}} \sum_j \mathcal{L}^{(j)}$ ,  
 $\tilde{\mathbf{y}} \leftarrow \tilde{\mathbf{y}} - \alpha \nabla_{\tilde{\mathbf{y}}} \sum_j \mathcal{L}^{(j)}$ , and  
 $\tilde{\eta} \leftarrow \tilde{\eta} - \alpha \nabla_{\tilde{\eta}} \sum_j \mathcal{L}^{(j)}$
  - 13: **end for**
  - 14: **for**  $i = 1$  to  $M$  **do**
  - 15:   Compute nearest embedding for every distilled word  
 $\tilde{\mathbf{x}}_i^* = \{\text{NearestEmbed}(\tilde{x}_{i,j})\}_{j=1}^s$
  - 16:   Decode embedding into text  
 $\tilde{\mathbf{z}}_i = \left\{ \text{WordFromEmbed}(\tilde{x}_{i,j}^*) \right\}_{j=1}^s$
  - 17: **end for**
  - 18:  $\tilde{\mathbf{z}} = \{\tilde{z}_i\}_{i=1}^M$
- Output:** distilled data  $\tilde{\mathbf{x}}$ ; distilled labels  $\tilde{\mathbf{y}}$ ; optimized learning rate  $\tilde{\eta}$ ; nearest sentences  $\tilde{\mathbf{z}}$
-

## 4.2 Image Data

The LeNet model we use with MNIST achieves nearly SOTA results, 99% accuracy, so it is sufficient to use distillation accuracy when describing distillation performance with it. However, AlexCifarNet only achieves 80% on CIFAR10 so it is helpful to use the 2 relative metrics when describing this set of distillation results.

**Baselines.** It is useful to compare dataset distillation against several other methods of dataset reduction. We use the following baselines suggested by Wang et al. (2018).

- **Random real images:** We randomly sample the same number of real images per class from the training data. These images are used for two baselines: training neural networks and training K-Nearest Neighbors classifiers.
- **Optimized real images:** We sample several sets of random real images as above, but now we choose the 20% of these sets that have the best performance on training data. These images are used for one baseline: training neural networks.
- **$k$ -means:** We use  $k$ -means to learn clusters for each class, and keep the resulting centroids. These images are used for two baselines: training neural networks and training K-Nearest Neighbors classifiers.
- **Average real images:** We compute the average image for each class and use it for training. These images are used for one baseline: training neural networks.

Each of these baseline methods produces a small set of images that can be used to train models. All four of the baseline methods are used to train and test LeNet and AlexCifarNet on their respective datasets. Additionally, two of the baseline methods are used to also train K-Nearest Neighbor classifiers to compare performance against neural networks. The results for these six baselines, as determined by Wang et al. (2018), are shown in Table 2.

**Fixed initialization.** When the network initialization is fixed between the distillation and training phases, synthetic images produced by dataset distillation result in high distillation accuracies. The SLDD algorithm produces images that result in equal or higher accuracies when compared to the original DD algorithm. For example, DD can produce 10 distilled images that train a LeNet model up to 93.76% accuracy on MNIST (Wang et al., 2018). Meanwhile, SLDD can produce 10 distilled images that train the same model up to 96.13% accuracy (Figure 1). The full distilled labels for these 10 images are laid out in Table 1. SLDD can even produce a tiny set of just 5 distilled images that train LeNet to 91.56% accuracy. As can be seen in Figure 6, the 90% distillation size (i.e. the minimum number of images needed to achieve 90% of the original accuracy) of MNIST with fixed initializations is  $d_A = 5$ , and while adding more distilled images typically increases distillation accuracy, this begins to plateau after five images. Similarly, SLDD provides a 7.5% increase in 100-sample distillation ratio (6% increase in distillation accuracy) on CIFAR10 over DD. Based on these results, detailed further in Table 2, it appears that SLDD is even more effective than DD at distilling image data into a small number of samples. This intuitively makes sense as the learnable labels used by SLDD increase the capacity of the distilled dataset for storing information.

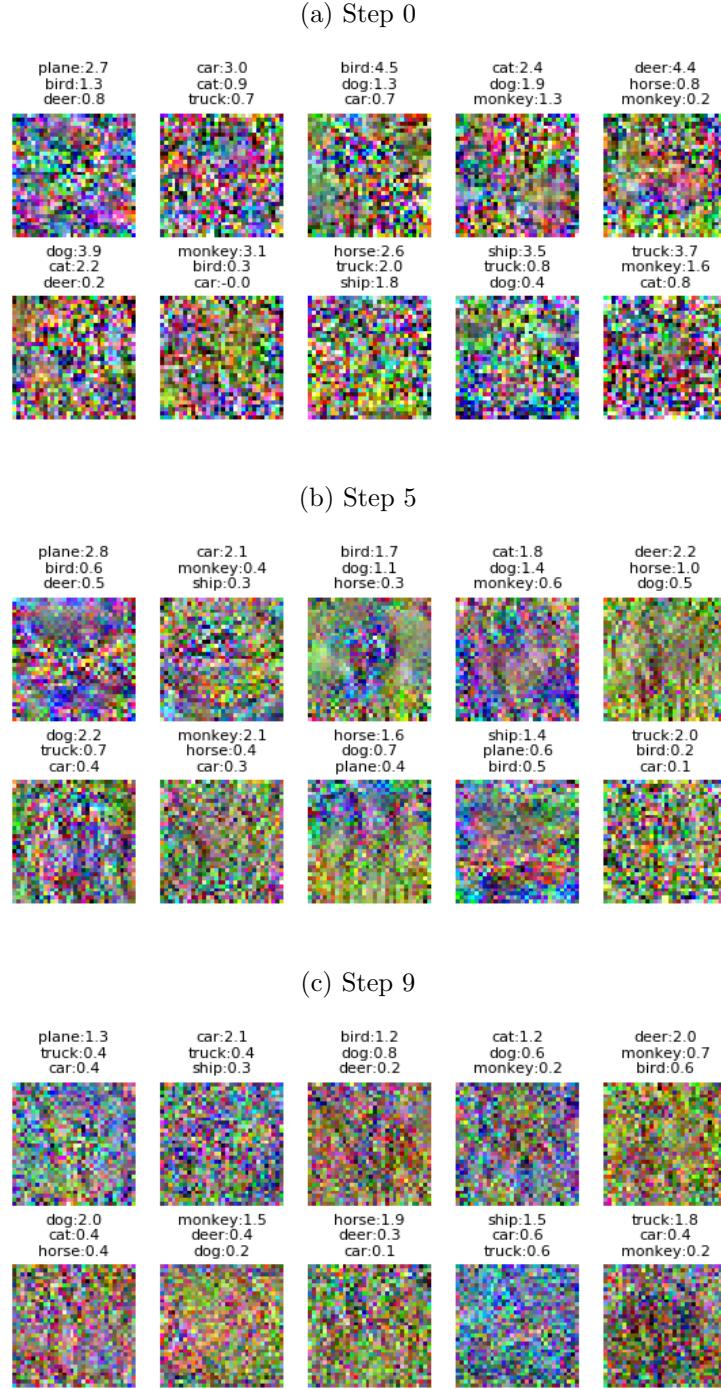


Figure 5: SLDD can learn 100 distilled CIFAR10 images that train networks with fixed initializations from 12.9% distillation accuracy to 60.0% ( $r_{100} = 75.0$ ). Each image is labeled with its top 3 classes and their associated logits. Only 3 of the 10 steps are shown.

**Random initialization.** It is also of interest to know whether distilled data are robust to network initialization. Specifically, we aim to identify if distilled samples store information only about the network initializations, or whether they can store information contained within the training data. To this end, we perform experiments by sampling random network initializations generated using the Xavier Initialization (Glorot and Bengio, 2010). The distilled images produced in this way are more representative of the training data but generally result in lower accuracies when models are trained on them. Once again, images distilled using SLDD lead to higher distillation accuracies than DD when the number of distilled images is held constant. For example, 100 MNIST images learned by DD result in accuracies of  $79.5 \pm 8.1\%$ , while 100 images learned by SLDD result in accuracies of  $82.75 \pm 2.75\%$ . There is similarly a 3.8% increase in 100-sample distillation ratio (3% increase in distillation accuracy) when using SLDD instead of DD on CIFAR10 using 100 distilled images each. These results are detailed in Table 2. It is also interesting to note that the actual distilled images, as seen in Figures 7 and 8, appear to have much clearer patterns emerging than in the fixed initialization case. These results suggest that DD, and even more so SLDD, can be generalized to work with random initializations and distill knowledge about the dataset itself when they are trained this way. All the mean and standard deviation results for random initializations in Table 2 are derived by testing with 200 randomly initialized networks.

### 4.3 Text Data

As mentioned above, TDD does not work in the space of the original raw data, but rather produces synthetic samples from the embedding space. Because each distilled ‘sentence’ is actually a matrix, the embedding layer is applied only to real sentences from the training data and not the matrices coming from the distilled data. For text experiments, we use the IMDB sentiment analysis dataset, the Stanford Sentiment Treebank 5-class task (SST5) (Socher et al., 2013), and the Text Retrieval Conference question classification tasks with 6 (TREC6) and 50 (TREC50) classes (Voorhees et al., 1999). The text experiments are performed with three different networks: a fairly shallow but wide CNN (TextConvNet), a bi-directional RNN (Bi-RNN) (Schuster and Paliwal, 1997), and a bi-directional Long Short-Term Memory network (Bi-LSTM) (Hochreiter and Schmidhuber, 1997). These models do not all achieve the same accuracies on the text datasets we use in distillation experiments, so the models’ original accuracies are detailed in Table 4. The distillation ratios appearing in this section are calculated based on these accuracies. The exact architectures of these models are detailed in the online appendix.

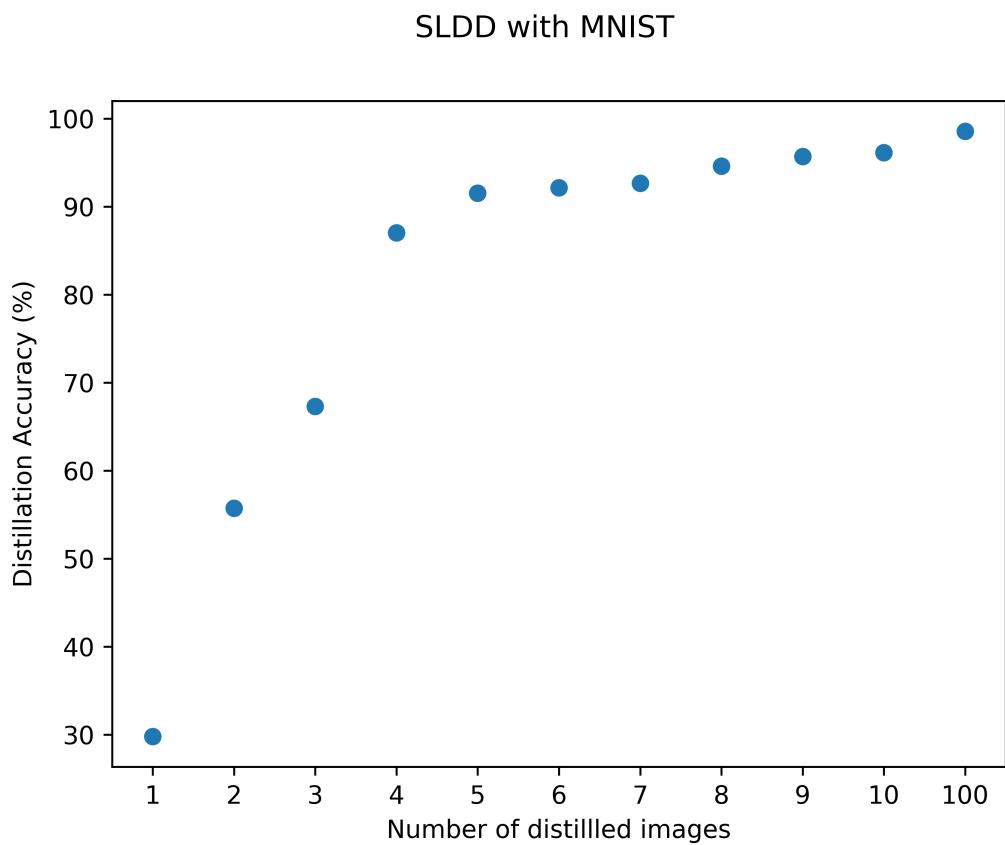


Figure 6: Distillation accuracy on MNIST with LeNet for different distilled dataset sizes.

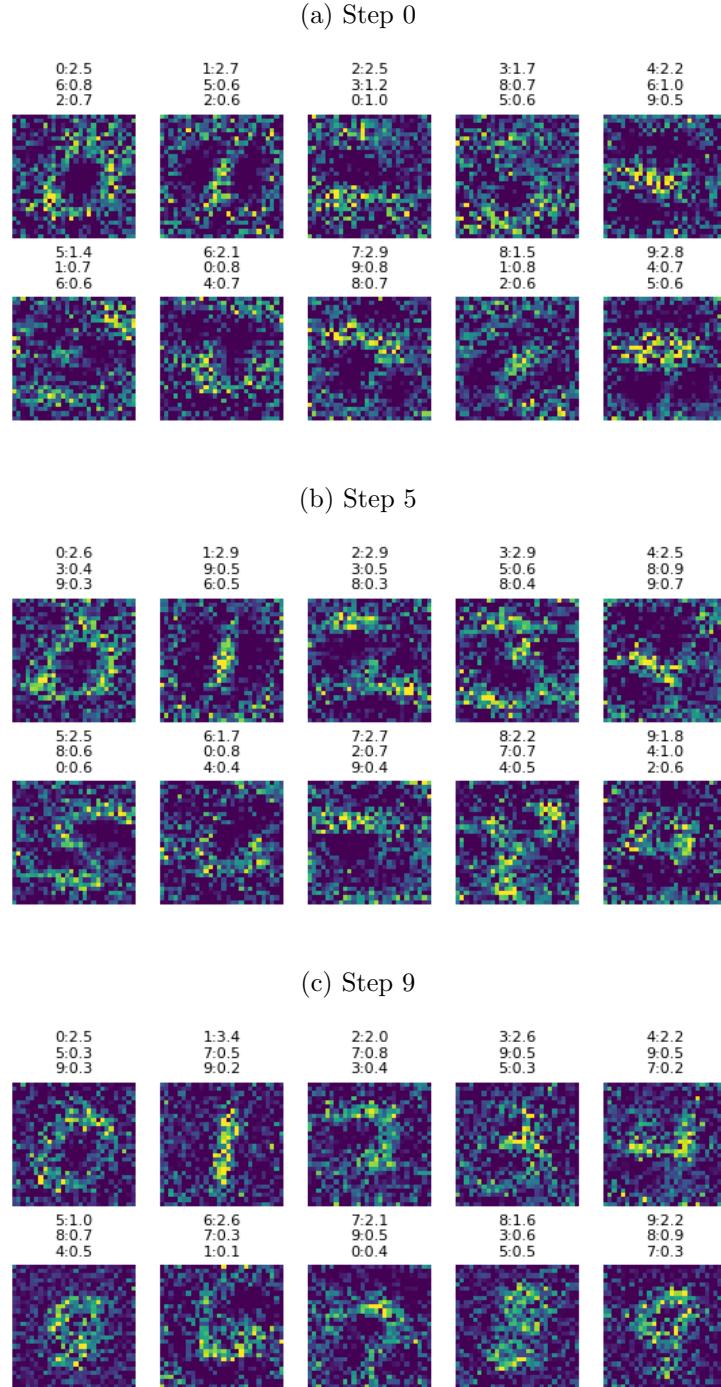


Figure 7: SLDD can learn 100 distilled MNIST images that train networks with random initializations from  $10.09\% \pm 2.54\%$  distillation accuracy to  $82.75\% \pm 2.75\%$  ( $r_{100} = 83.6$ ). Each image is labeled with its top 3 classes and their associated logits. Only 3 of the 10 steps are shown.

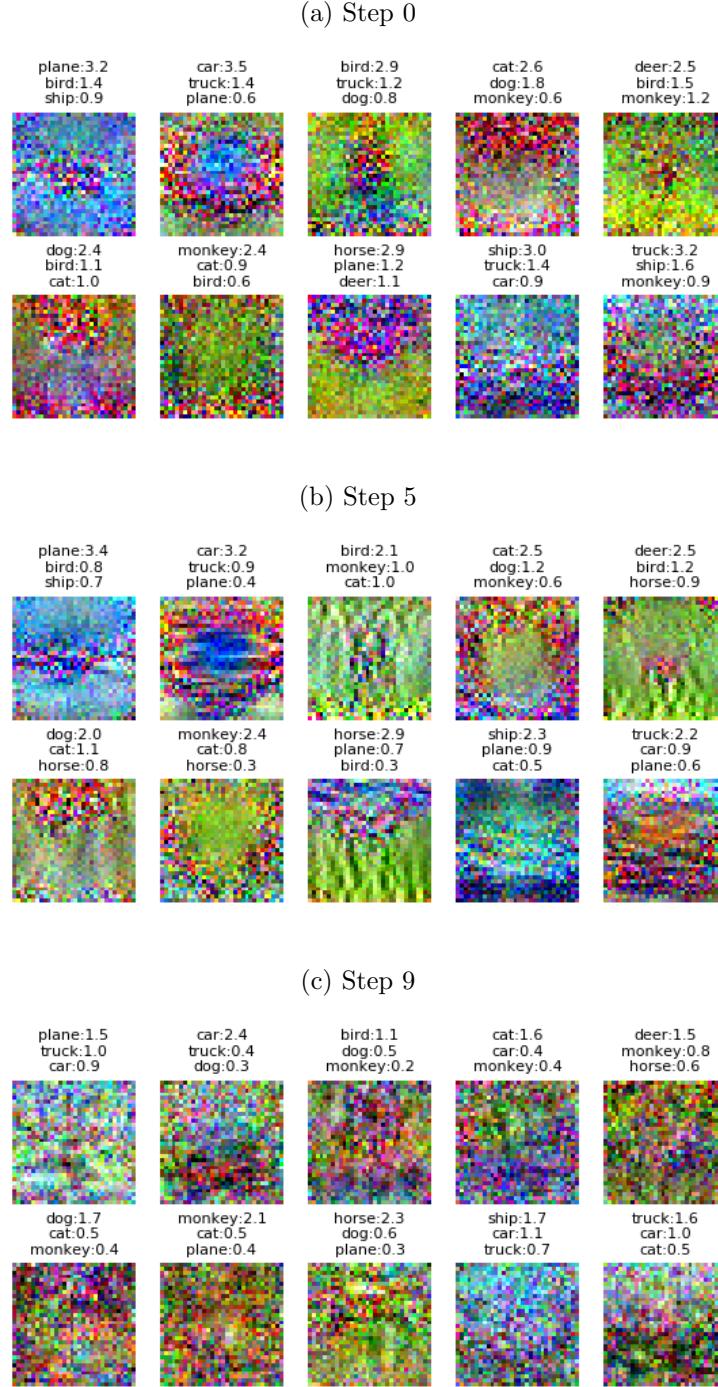


Figure 8: SLDD can learn 100 distilled CIFAR10 images that train networks with random initializations from  $10.17\% \pm 1.23\%$  distillation accuracy to  $39.82\% \pm 0.83\%$  ( $r_{100} = 49.8$ ). Each image is labeled with its top 3 classes and their associated logits. Only 3 of the 10 steps are shown.

Table 2: Means and standard deviations of distillation and baseline accuracies on image data. All values are percentages. The first four baselines are used to train the same neural network as in the distillation experiments. The last two baselines are used to train a K-Nearest Neighbors classifier. Experiments with random initializations have their results listed in the form [mean  $\pm$  standard deviation] and are based on the resulting performance of 200 randomly initialized networks.

	SLDD accuracy			DD accuracy			Used as training data in same # of GD steps			Used in K-NN			
	Fixed	Random	Fixed	Random	Rand.	real	Optim.	real	$k$ -means	Avg.	real	Rand.	real
MNIST	<b>98.6</b>	82.7 $\pm$ 2.8	96.6	79.5 $\pm$ 8.1	68.6 $\pm$ 9.8	73.0 $\pm$ 7.6	76.4 $\pm$ 9.5	77.1 $\pm$ 2.7	71.5 $\pm$ 2.1	<b>92.2</b> $\pm$ <b>0.1</b>			
CIFAR10	<b>60.0</b>	<b>39.8</b> $\pm$ <b>0.8</b>	54.0	36.8 $\pm$ 1.2	21.3 $\pm$ 1.5	23.4 $\pm$ 1.3	22.5 $\pm$ 3.1	22.3 $\pm$ 0.3	18.8 $\pm$ 1.3	29.4 $\pm$ 0.3			

Table 3: Means and standard deviations of TDD and baseline accuracies on text data using TextConvNet. All values are percentages. The first four baselines are used to train the same neural network as in the distillation experiments. The last two baselines are used to train a K-Nearest Neighbors classifier. Each result uses 10 GD steps aside from IMDB with  $k$ -means and TREC50 which had to be done with 2 GD steps due to GPU memory constraints and also insufficient training samples for some classes in TREC50. The second TREC50 row uses TDD with 5 GD steps with 4 images per class. Experiments with random initializations have their results listed in the form [mean  $\pm$  standard deviation] and are based on the resulting performance of 200 randomly initialized networks.

	TDD accuracy			Used as training data in 10 GD steps			Used in K-NN					
	Fixed	Random	Rand.	real	Optim.	real	$k$ -means	Avg.	real	Rand.	real	$k$ -means
IMDB	<b>75.0</b>	<b>73.4</b> $\pm$ <b>3.3</b>	49.7 $\pm$ 0.9	49.9 $\pm$ 0.8	49.9 $\pm$ 0.6	50.0 $\pm$ 0.1	50.0 $\pm$ 0.1	50.0	0.1	50.0	0.1	50.0 $\pm$ 0.0
SST5	<b>37.5</b>	<b>36.3</b> $\pm$ <b>1.5</b>	21.2 $\pm$ 4.9	24.6 $\pm$ 2.6	19.6 $\pm$ 4.5	21.3 $\pm$ 4.1	23.1 $\pm$ 0.0	20.9	0.0	20.9	0.0	20.9 $\pm$ 2.1
TREC6	<b>79.2</b>	<b>77.3</b> $\pm$ <b>2.9</b>	37.5 $\pm$ 10.1	44.6 $\pm$ 7.5	34.4 $\pm$ 13.0	28.0 $\pm$ 9.5	31.5 $\pm$ 9.9	50.5	6.8			
TREC50	<b>57.6</b>	11.0 $\pm$ 0.0	8.2 $\pm$ 6.0	9.9 $\pm$ 6.6	14.7 $\pm$ 5.5	12.5 $\pm$ 6.4	15.4 $\pm$ 5.1	<b>45.1</b> $\pm$ <b>6.6</b>				
TREC50 <sup>2</sup>	67.4	42.1 $\pm$ 2.1										

Table 4: Model accuracies when trained on full text datasets.

Model	Dataset	# of Classes	Accuracy
TextConvNet	IMDB	2	87.1%
Bi-RNN	SST5	5	41.0%
Bi-LSTM	TREC6	6	89.4%
TextConvNet	TREC50	50	84.4%

Table 5: Distillation ratios for text datasets and their associated neural networks. Experiments with random initializations have their results listed in the form [mean  $\pm$  standard deviation] and are based on the resulting performance of 200 randomly initialized networks.

Model	Dataset	Number of Distilled Sentences (M)	Distillation Ratio ( $r_M$ )	
			Fixed	Random (Mean $\pm$ SD)
TextConvNet	IMDB	2	89.9	80.0 $\pm$ 6.3
TextConvNet	IMDB	20	91.5	85.2 $\pm$ 3.2
Bi-RNN	SST5	5	87.7	57.0 $\pm$ 5.7
Bi-RNN	SST5	100	89.8	66.8 $\pm$ 5.4
Bi-LSTM	TREC6	6	97.8	69.3 $\pm$ 9.8
Bi-LSTM	TREC6	120	98.2	78.9 $\pm$ 6.3
TextConvNet	TREC50	500	57.6	11.0 $\pm$ 0.0
TextConvNet	TREC50	1000	67.4	42.1 $\pm$ 2.1

**Baselines.** We consider the same six baselines as in the image case but modify them slightly so that they work with text data.

- **Random real sentences:** We randomly sample the same number of real sentences per class, pad/truncate them, and look up their embeddings. These sentences are used for two baselines: training neural networks and training K-Nearest Neighbors classifiers.
- **Optimized real sentences:** We sample and pre-process different sets of random real sentences as above, but now we choose the 20% of the sets that have the best performance. These sentences are used for one baseline: training neural networks.
- **$k$ -means:** First, we pre-process the sentences. Then, we use  $k$ -means to learn clusters for each class, and use the resulting centroids to train. These sentences are used for two baselines: training neural networks and training K-Nearest Neighbors classifiers.
- **Average real sentences:** First, we pre-process the sentences. Then, we compute the average embedded matrix for each class and use it for training. These sentences are used for one baseline: training neural networks.

Each of these baseline methods produces a small set of sentences, or sentence embeddings, that can be used to train models. All four of the baseline methods are used to train and test the TextConvNet on each of the text datasets. Additionally, two of the baseline methods are used to also train K-Nearest Neighbor classifiers to compare performance against neural networks. The baseline results are shown in Table 3.

**Fixed initialization.** When the network initialization is fixed between the distillation and training phases, synthetic text produced by text dataset distillation also results in high model accuracies. For example, TDD can produce 2 distilled sentences that train the TextConvNet up to a distillation ratio of 89.88% on the IMDB dataset. Even for far more difficult language tasks, TDD still has impressive results but with larger distilled datasets. For example, for the 50-class TREC50 task, it can produce 1000 distilled sentences that train the TextConvNet to a distillation ratio of 79.86%. Some examples of TDD performance are detailed in Table 3 and Table 5. The distilled text embeddings from the six-sentence Trec6 experiments are visualized in Figure 9. However, since these distilled text embeddings are still in the GloVe embedding space, it may be difficult to interpret them visually. We provide a more natural method for analyzing distilled sentences by using nearest-word decoding to reverse the GloVe embedding. We find the nearest word to each distilled vector based on Euclidean distances. The result of this decoding is an approximation of the distilled sentence in the original text space. We list the decoded distilled sentences corresponding to the matrices from Figure 9 in Table 6, along with their respective label distributions. These sentences can contain any tokens found in the TREC6 dataset, including punctuation, numbers, abbreviations, etc. The sentences do not have much overlap. This is consistent with the distilled labels which suggest that each sentence corresponds strongly to a different class. It appears that the TDD algorithm encourages the separation of classes, at least when there are enough distilled samples to have one or more per class. Additional results and visualizations for TDD with fixed initialization can be found in the online appendix.

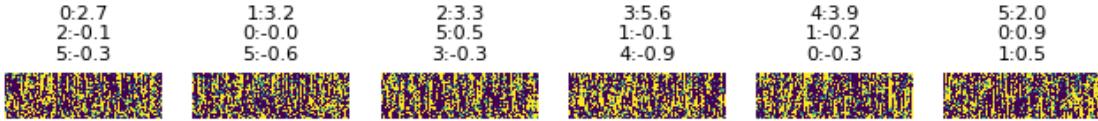


Figure 9: TDD can learn 6 distilled sentences of length 30 that train networks with fixed initializations from 12.6% to 87.4% ( $r_6 = 97.8$ ). Each image corresponds to a distilled embedding and is labeled with its top 3 classes and their associated logits.

Distilled Sentence	Label Class					
	0	1	2	3	4	5
allan milk banned yellow planted successfully introduced bombay 1936 grass mines iron delhi 1942 male heir throne oath clouds 7th occur millennium smoking flows truth powder judiciary pact slim profit	<b>2.72</b>	-0.48	-0.07	-0.62	-0.53	-0.27
whom engineer grandfather joan officer entered victoria 1940s taxi romania motorcycle italian businessman photographer powerful driving u brilliant affect princess 1940s enemies conflicts southwestern retired cola appearances super dow consumption	-0.05	<b>3.21</b>	-1.15	-0.79	-0.71	-0.64
necessarily factors pronounced pronounced define bow destroying belonged balls 1923 storms buildings 1925 victorian sank dragged reputation sailed nn occurs darkness blockade residence traveled banner chef ruth rick lion psychology	-0.67	-0.66	<b>3.28</b>	-0.27	-0.77	0.47
accommodate accommodate peak 2.5 adults thin teenagers hike aged nurse policeman admit aged median philippines define baghdad libya ambassador admit baseman burma inning bills trillion donor fined visited stationed clean	-0.98	-0.14	-1.12	<b>5.57</b>	-0.85	-1.85
suburb ports adjacent mountains nearest compare hilton volcano igor nebraska correspondent 1926 suburb sailed hampshire hampshire gathering lesson proposition metric copy carroll sacred moral lottery whatever fix o completed ultimate	-0.29	-0.22	-0.36	-1.01	<b>3.86</b>	-0.44
advertising racism excuse d nancy solved continuing congo diameter oxygen accommodate provider commercials spread pregnancy mideast ghana attraction volleyball zones kills partner serves serves congressman advisory displays ranges profit evil	0.94	0.53	-0.85	0.08	-0.83	<b>1.99</b>

Table 6: Nearest sentence decodings corresponding to the distilled embeddings in Figure 9. Each sentence is accompanied by the logits associated with each value of its distilled label. The top class for each distilled sentence is bolded.

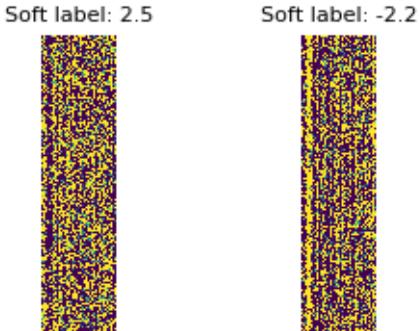


Figure 10: TDD can learn 2 distilled sentences of length 400 that train networks with random initializations from 50.0% to  $69.6\% \pm 5.5\%$  ( $r_2 = 79.96$ ). Each image corresponds to a distilled embedding and is labeled with its soft label value. The soft label is a scalar as this is a binary classification task.

**Random initialization.** When using random initialization the performance decrease for TDD is similar to that for SLDD. TDD can produce two distilled sentences that train the TextConvNet with random initialization up to a distillation ratio of 79.96% on IMDB, or 20 distilled sentences that train it up to a distillation ratio of 85.22%. This is only slightly lower performance than in the fixed initialization case. However, there is a larger difference in performance between fixed and random initializations for the recurrent networks. For TREC6, TDD can produce six distilled sentences that train a randomly initialized Bi-LSTM to a distillation ratio of 69.33%, or 120 distilled sentences that train it to a distillation ratio of 78.87%. All the mean and standard deviation results for random initializations in Table 3 and Table 5 are derived by testing with 200 randomly initialized networks. The distilled text embeddings from the IMDB experiment with two distilled sentences are visualized in Figure 10. We list the decoded distilled sentences corresponding to these matrices in Table 7. These sentences can contain any tokens found in the IMDB dataset, including punctuation, numbers, abbreviations, etc. Since this is a binary sentiment classification task, each label is a scalar. If a probability is needed, a sigmoid function can be applied to the scalar soft labels. In this case, the distillation algorithm appears to have produced one sentence with a positive associated sentiment, and one with a negative sentiment. Curiously, the model appears to have overcome the challenge of having to describe these long sentences with a single scalar by using duplication. For example, in the second sentence, corresponding to the negative label, negative words like ‘dump’, ‘stupid’, and ‘shoddy’ are all repeated several times. In such a way, the algorithm is likely assigning lower sentiment scores to these words than other ones, while using only a single label. Additional results for TDD with random initialization can be found in the online appendix.

## 5. Conclusion

By introducing learnable distilled labels we have increased distillation accuracy across multiple datasets by up to 6%. By enabling text distillation, we have also greatly increased the types of datasets and architectures with which distillation can be used.

Distilled Sentence	Label
editor panoramic brewing swat regency medley arts fleet attained hilary novelist rugged medal who abbot has sweden new ensemble member understands alba archaeologist operatic intercontinental martian marshal paste smooth titular english-language adaptation songwriter historian enlightenment royal gaelic ceo author macarthur skipper honored excellence distance most endings collaborations his mythological polanski mayer choreographer grisham eminent brooke sympathies modelling vitality dictionary dedication farewell enjoys energetic jordan equality lectures sophia elijah maureen novelist sanjay zeta blues colors talent dylan dominic anne 2003 who pavarotti dedication illustration mary exhibited scenic award colonial filmography previews dialect scenery steely representative adventures opens favourites macarthur economic erected mabel biography def waterfront awards tenor artists coin choreographer art securing anna earned helen emmy vibrant repertoire witty prestigious marquis classic splendor kindness lyrical architecture climbed princess quotes ava touring eponymous johnson strengths title county acclaimed courage revolt prolific classic honors rocker unforgettable ageing founding vigorous renowned deepest operatic female intertwined impressionist poignant vanilla impassioned esp isle camaraderie romania regal casablanca caribbean friendship friendship todd meadows hosted billie regina intimate poetic digital warmth savvy great symphony twilight thrills meditation ebert emperor snail remarkable rhys classics talent kristin patricia kim ballads climber jordan advancing phil i. robbie beatrice brighter pacifist hebrew	2.53
que yeah dump misled speculate bait substandard uncovered lump corpses 911 punched whopping discarded ref dollar were dough sided brink unconscious tomatoes locking trash burying punched diversion grenade overboard cashier wards agreeing brent prematurely knife randomly stupid buster wipe virus pap waste inflated loan patient peg eliminates nudity worms ?? rotten shoddy strangled substandard boil whopping tunnels steep unjust dummy satisfactory mistakenly adopt tightened bloated hacked misled dump 3-4 untrue contaminated bureaucracy waste chopping bait gram rotting intentional washed unhealthy liar corpses hospitals ?? filthy lifts freeze cabin contaminated shutting rents discrimination frying stomach toxic piles riddled negligible shattering fewer radioactive melt censors gram spilling housing saif subway sabotage booby nuke bb wasteland nuisance knock false useless gonna budget useless grenades cigarettes financed mandatory mine parked bb !! hangover riddled hijack shut cheaply unused sim mole trailers collide misunderstanding lump projected litter fx worst tamil rotten sewer substandard financing leak ?? gonna any claiming tails giancarlo explode frustrating flooding slit nixon prematurely pay leftover chimney outs shotgun stupid surplus minimum shutting punched radioactive stupid worthless worst shoddy nudity rotting workable worthless doses hanged rub be busted stupid faulty lame gunshot grenade dump trash wrist alastair hurry suffice detect rigged transformers unrated ; deplorable scrape hacked	-2.21

Table 7: Nearest sentence decodings corresponding to the distilled embeddings in Figure 10. Each sentence is accompanied by its associated soft label. Only the first 200 (out of 400) words are shown for each sentence.

However, even with SLDD and TDD, there are still some limitations to dataset distillation. The network initializations used for both SLDD and TDD all come from the same distribution, and no testing has yet been done on whether a single distilled dataset can be used to train networks with different architectures. Further investigations are needed to determine more precisely how well dataset distillation can be generalized to work with more variation in initializations, and even across networks with different architectures.

Interestingly, the initialization of distilled labels appears to affect the performance of dataset distillation. Initializing the distilled labels with ‘hard’ label values leads to better performance than with random initialization, possibly because it encourages class separation earlier on in the distillation process. However, it is not immediately clear whether it is better to separate similar classes (e.g. ‘3’ and ‘8’ in MNIST), thereby increasing the network’s ability to discern between them, or to instead keep those classes together, thereby allowing soft-label information to be shared between them. It may be interesting to explore the dynamics of the distillation process when using a variety of label initialization methods.

We have shown dataset distillation works with CNNs, bi-directional RNNs, and LSTMs. There is nothing in the dataset distillation algorithm that would limit it to these network types. As long as a network has a twice-differentiable loss function and the gradient can be back-propagated all the way to the inputs, then that network is compatible with dataset distillation.

Another promising direction is to use distilled datasets for speeding up Neural Architecture Search and other very compute-intensive meta-algorithms. If distilled datasets are a good proxy for performance evaluation, they can reduce search times by multiple orders of magnitude. In general, dataset distillation is an exciting new branch of knowledge distillation; improvements may help us not only better understand our datasets but also enable several applications related to efficient machine learning.

## Acknowledgments

We would like to thank Dr. Sebastian Fischmeister for providing us with the computational resources that enabled us to perform many of the experiments found in this work.

## References

- Anelia Angelova, Yaser Abu-Mostafam, and Pietro Perona. Pruning training sets for learning of object categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 494–501. IEEE, 2005.
- Olivier Bachem, Mario Lucic, and Andreas Krause. Practical coresnet constructions for machine learning. *arXiv preprint arXiv:1703.06476*, 2017.
- Cristian Bucilu, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 535–541. ACM, 2006.
- Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797. IEEE, 2018.
- David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional Transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Neamat El Gayar, Friedhelm Schwenker, and Günther Palm. A study of the robustness of knn classifiers trained using soft labels. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 67–80. Springer, 2006.
- R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x. URL <https://onlinelibrary.wiley.com/doi/10.1111/j.1469-1809.1936.tb02137.x>.
- Salvador Garcia, Joaquin Derrac, Jose Cano, and Francisco Herrera. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE transactions on pattern analysis and machine intelligence*, 34(3):417–435, 2012.
- Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256. PMLR, 2010.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680. Curran Associates, Inc., 2014.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

- Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*, 2016.
- Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. URL <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>.
- Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.
- Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. *Object Recognition with Gradient-Based Learning*, pages 319–345. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999. ISBN 978-3-540-46805-9. doi: 10.1007/3-540-46805-6\_19. URL [https://doi.org/10.1007/3-540-46805-6\\_19](https://doi.org/10.1007/3-540-46805-6_19).
- Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017.
- Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.
- Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. *arXiv preprint arXiv:1603.01354*, 2016.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.
- Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

- Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.
- Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.
- Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, 2013.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243*, 2019.
- Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2(Nov):45–66, 2001.
- Isaac Triguero, Joaquín Derrac, Salvador García, and Francisco Herrera. A taxonomy and experimental study on prototype generation for nearest neighbor classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(1):86–100, 2011.
- Ivor W Tsang, James T Kwok, and Pak-Ming Cheung. Core vector machines: Fast SVM training on very large data sets. *Journal of Machine Learning Research*, 6(Apr):363–392, 2005.
- Ellen M Voorhees et al. The TREC-8 question answering track report. In *the Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, volume 99, pages 77–82, 1999.
- Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *International Conference on Machine Learning*, pages 1058–1066, 2013.
- Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.
- L Yu, W Zhang, J Wang, and Y Yu. SeqGAN: sequence generative adversarial nets with policy gradient. In *AAAI-17: Thirty-First AAAI Conference on Artificial Intelligence*, volume 31, pages 2852–2858. Association for the Advancement of Artificial Intelligence (AAAI), 2017.